DOCUMENT RESUME

ED 384 638 TM 023 804

AUTHOR O'Sullivan, Christine
TITLE The Cost of Performance Assessment in Science: The NAEP Perspective.
INSTITUTION Educational Testing Service, Princeton, N.J.
SPONS AGENCY National Center for Education Statistics (ED), Washington, DC.
PUB DATE Apr 95
NOTE 14p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 19-21, 1995).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Cost Estimates; *Educational Assessment; Elementary School Students; Elementary Secondary Education; *Field Tests; Grade 4; Grade 8; Grade 12; High School Students; National Surveys; *Science Tests; *Test Construction; Test Results
IDENTIFIERS Hands on Science; *National Assessment of Educational Progress; Paper and Pencil Tests; *Performance Based Evaluation

ABSTRACT

The cost of including hands-on tasks in the National Assessment of Educational Progress (NAEP) science assessment was analyzed. In recognition of growing interest in performance assessment in science, the framework for the NAEP science assessment called for the inclusion of hands-on tasks with hands-on activities and questions about the activities. The analysis considered the incremental costs of assessing the framework with a combination of paper-and-pencil blocks and hands-on task blocks over and above the costs of assessing the framework by paper-and-pencil alone. A field test was administered to 560 students at grade 4, 525 at grade 8, and 435 at grade 12. The time to develop a hands-on task block, with associated professional staff time, was found to be very similar to the time taken to develop an equivalent set of paper-and-pencil questions, and scoring was determined to take about the same amount of time. However, additional costs need to be added for small-scale review, pilot testing, dealing with test kit manufacturers, quality control of kits, and repacking. Nine tables present details of the analyses. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 384 638

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
CHRISTINE O'SULLIVAN

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

THE COST OF PERFORMANCE ASSESSMENT IN SCIENCE: THE NAEP PERSPECTIVE

Christine O'Sullivan
Educational Testing Service

BEST COPY AVAILABLE

Paper presented as part of the symposium "The Cost of Performance Assessment in Science" at the 1995 annual meeting of the National Council of Measurement in Education in San Francisco, California, April, 1995. This work was funded by the National Center for Education Statistics.

TM023804

THE COST OF PERFORMANCE ASSESSMENT IN SCIENCE
THE NAEP PERSPECTIVE[1]

Background Information

The National Assessment of Educational Progress (NAEP) is a congressionally mandated project of the National Center for Education Statistics, the U.S. Department of Education. The Commissioner of Education Statistics is responsible, by law, for carrying out the NAEP projects through competitive awards to qualified organizations. NAEP reports directly to the Commissioner. The Commissioner is also responsible for providing continuing reviews, including validation studies and solicitation of public comment, on NAEP's conduct and usefulness.

NAEP is the government's primary indicator of what students at grades 4, 8, and 12 know and can do in different subject areas. In addition to questions in specific subject areas, NAEP also gathers background data on student, teacher, and school variables that may be related to subject-area performance.

NAEP is designed to

- report information on student performance, including trends over time, for the nation, regions, males and females, racial/ethnic groups, and types of communities;
- report information collected from students about variables such as homework, television viewing, reading materials in the home, and parental education;
- present results in terms of proficiency scales that describe the particular kinds of tasks students can perform in a given subject;
- report information collected from students and their teachers regarding classroom experiences, instructional techniques, resources ranging from computers to library usage, and students' exposure to particular topics;
- serve as a national information system, from which states can voluntarily seek comparative data.

---

[1]The author wishes to thank the following people for the information and help they provided: Hilary Persky, Stephen Lazer, Anne Stauffer, and Mario Yepes-Baraya from ETS; Sandy Rieder from Westat, Inc.; and Brent Studer, Brad Thayer, Tim Robinson, and Patrick Bourgeacq from National Computer Systems.

NAEP is not designed to

- report individual student results;
- promote particular course offerings or instructional approaches (the assessments measure outcomes, notwithstanding students' educational experiences).

Purposes of the NAEP Analysis

The purpose of the NAEP analysis is to tease out the incremental costs of assessing the framework with a combination of paper-and-pencil blocks and hands-on task blocks over and above the costs of assessing the framework by paper-and-pencil alone.

Definitions and Procedures

The tasks involved in the preparation, administration, and scoring of the 1993 NAEP science field test were undertaken by four major organizations (Table 1). The Council of Chief State School Officers (CCSSO) oversaw the development of the assessment framework and specifications; Educational Testing Service (ETS) developed the assessment, using the framework and specifications as a blueprint; Westat, Inc. carried out the student sampling and administered the assessment; and National Computer Systems (NCS) printed the books, shipped and received supplies for the assessment, and scored the student responses.

Table 1: Areas of Responsibility

ORGANIZATION	TASK
CCSSO	Framework Development
ETS	Assessment Development
Westat, Inc.	Sampling and Administration
NCS	Printing, Shipping, Receiving, and Scoring

The organizations involved in the NAEP science assessment are situated in different areas of the country, so estimates of cost are based on number of hours, except for the cost of hands-on kits, which is given in dollars. Costs that have been considered in the analysis are those that were incurred during development of tasks, training of field supervisors, quality control, repacking, and scoring. Some costs could not be factored in because of the difficulty of separating out the steps directly related to the hands-on task section of the assessment. Other costs, such as preparation of final copy and printing and shipping of test booklets, were not considered because, if hands-on tasks had not been used, additional blocks of paper-and-pencil questions would most likely have been administered.

OVERVIEW OF ACTIVITIES

NAEP Framework and Specifications

The framework and specifications for the NAEP science assessment were developed in 1991-1992 under the auspices of the National Assessment Governing Board through a consensus process managed by the Council of Chief State School Officers[2]. The questions for the assessment were field tested in 1993. Based on the results of the field test, the science assessment was assembled for administration in 1994. The assessment, however, was postponed and will be administered in 1996.

The framework developers recognized that there had been a growth of support within the science education community for an increased use of performance-based tasks as a means of assessing student proficiency in science. It was felt that hands-on activities could measure aspects of student's science skills and understanding not probed by more traditional paper-and-pencil tests. Thus, the framework for the NAEP science assessment called for the inclusion of performance tasks (hands-on tasks) consisting of hands-on activities and sets of questions related to the activities for which students provide written responses.

Hands-on Assessment Tasks

The hands-on tasks were constructed on the basis of descriptions given in the NAEP framework. They were written by members of ETS staff as well as by outside individuals. A general shell for all tasks was developed and task developers worked within this shell. Specifically, each task contains introductory information explaining briefly what is to be accomplished in the task; a diagram of the materials and equipment in the kit accompanying the task; directions for accomplishing the task; and multiple-choice and constructed-response questions relating to the task. While this shell is standard, there are differences between tasks in the way the directions and questions are presented. Some tasks are designed so that

---

[2] Science Framework for the 1994 National Assessment of Educational Progress (pre-publication draft); The National Assessment Governing Board, U.S. Department of Education.

students are prompted to write down responses as they proceed through the task. In other tasks the students have to complete the task and then answer the questions. A third type of task presents a problem that students have to solve given certain materials and equipment. Each task, together with its related questions, makes up a "block" for assessment purposes.

Seventeen tasks were selected for field testing, of which four were intended for students in more than one grade. All had undergone rigorous pretesting. Each hands-on task and its accompanying questions were first subjected to a small-scale review in which approximately six students performed the tasks. Kits that had been put together in-house were used for this review. After completing a task, the students were interviewed to discover any problems they may have encountered while performing the task. The interview responses were used as a basis for improving the clarity of the instructions and questions, and also to improve the quality and types of equipment. The revised tasks were then piloted in approximately fifteen local schools using prototype kits supplied by two kit manufacturers. Students were watched closely as they performed each task and were then interviewed. The interview responses were used as a basis for further revision.

Field Test Chronology

The field-test process followed the timeline described in Table 2.

Table 2: NAEP Field Test Schedule

ACTIVITY	DATES
Task Development	April - July, 1992
Equipment Preparation	April - November, 1992
Booklet Printing	October - November, 1992
Field-test Administration	January - February, 1993
Scoring	February - March, 1993

Student Sample

Table 3 shows the total number of paper-and-pencil blocks and hands-on task blocks, booklets, and student sample size for the field test. Each booklet was administered to 560 students at grade 4, 525 students at grade 8, and 435 students at grade 12. There were approximately 160 science sessions at each grade level and 80 sessions for a grade 12 special study. These sessions were held in approximately 120 schools at each of grades 4, 8, and 12. Intact classes containing approximately 30 students were used for each session.

Table 3: Number of Blocks, Booklets, and Sample Size for the 1993 Science Field Test

Grade	Number of Paper-and-pencil Blocks	Number of Hands-on Task Blocks	Booklets	Sample (Approx.)
4	14	7	7	3,500
8	14	7	7	3,500
12	20*	7	10	5,000
Total	48	21	24	12,000

* These include a grade 12 special study.

Administration

The assessment was administered by field supervisors employed by Westat, Inc. It was important that each student participating in the assessment be given exactly the same instructions, so written scripts were used during administration. The field supervisors were trained on the hands-on tasks by ETS staff. NCS sent the administration materials to the field supervisors, who then returned them to NCS for scoring.

Each sampled student took 3 blocks of cognitive items: 2 paper-and-pencil blocks and 1 hands-on task block. Grade 4 students were given 20 minutes to complete each block of questions and grade 8 and grade 12 students were given 30 minutes to complete each block of questions. Thus the testing time was 60 minutes for grade 4 students and 90 minutes for grade 8 and grade 12 students. In addition to the assessment, students were also asked to complete background questionnaires.

Scoring

Student responses were scored over a seven-week period, from February, 1993 to March, 1993 at NCS in Iowa City. A total of 450 constructed-response questions were scored, of which approximately 140 were in the hands-on task blocks. Approximately five hundred students answered each constructed-response question except for cross-grade questions, which were answered by about 1,000 students each.

Scoring guides for each question had been prepared during the development of questions and tasks and were subject to revision in light of student responses. The scoring

guides were analytical in nature and usually consisted of a maximum of 3 or 4 points per question. Prior to scoring, training packets containing student responses were assembled. These training packets consisted of anchor papers, practice papers, and qualification papers. The assembly of training packets took 7 days and involved staff from both ETS and NCS.

Four tables of ten scorers took seven weeks to score the questions. Each table had a trainer from ETS who was responsible for training the scorers and a table leader from NCS who was responsible for back-reading and making certain that the scorers were following the scoring guides. The constructed-response questions in each block were trained as a unit, after which the scorers proceeded to score each student's block of questions.

RESOURCES AND COSTS

The cost of including hands-on tasks in the NAEP science assessment is complicated by the fact that the additional costs to NAEP are the steps that are not usually followed for the paper-and-pencil blocks: namely, pilot testing, consultation with kit manufacturers, cost of kits, training of supervisors on use of kits, quality control of kits, boxing and shipping of kits to supervisors, and receiving and sorting kits after the assessment administration.

Steps that are common to both paper-and-pencil blocks and hands-on task blocks have been included for completeness. In the final cost analysis, however, these common steps have been excluded.

Costs are given in terms of professional staff hours. This allows for regional variations in pay. In addition, the average cost of the kits used in the small-scale review and field test are given in dollars.

Task Development

Task development costs in terms of hours are summarized in Table 4. Some tasks were developed externally and some internally. The hours are given for those that were developed internally. An Instrument Development Committee made up of teacher and teacher educators was involved at all stages of NAEP assessment development. The hours taken for them to review a task are given.

Table 4: Number of Hours to Develop a Hands-on Task

TASK PREPARATION	NUMBER OF PEOPLE	HOURS	PEOPLE x HOURS
Draft 1	1	8	8
Review 1	4	2	8
Respond To Review	1	5	5
Draft 2	1	2	2
*Small-scale Review (Kit Assembly - 6)	1	4	4
*Small-scale Review	2	1	2
*Draft 3	1	2	2
*Pilot Test (2 Schools per Task)	2	1	2
*Analysis of Pilot Test	1	2	2
*Draft 4	1	2	2
Scoring Guide	1	16	16
Committee Review	4	3	12
Final Task	1	4	4
Administrative Support	1	25	25
TOTAL HOURS			94

* Steps unique to hands-on task block

Kit Preparation

The kits used in the small-scale review were assembled in-house. Parts were either bought locally or borrowed. Therefore, the cost per kit is difficult to calculate. Parts were also recycled. A fair estimate of the cost is probably about $2 a kit. The kits used in the pilot study and in the field test were prepared by two companies. Eight hundred and fifty of each of 17 kits were produced. The average cost per kit was approximately $6.

A number of steps have to be factored into the time taken by ETS staff to work with kit manufacturers. Specifications for kits had to be drawn up. Meetings were held with those kit manufacturers who had submitted bids. Staff members were consulted on numerous occasions, and a number of quality controls were carried out. Since there were 17 tasks, a total time was estimated and divided by the number of tasks (Table 5).

Table 5: Kit Preparation in ETS Staff Hours

ACTIVITY	HOURS/KIT
Specification Preparation	2
Meetings with Manufacturers	1
Consultations	3
Quality Control	2
TOTAL	8

Shipping and Receiving

NCS was responsible for receiving the kits from the manufacturers and sending them to field supervisors. It had been anticipated that approximately 10% of the kits would be checked, however, all of the kits in approximately half of the tasks were examined, and 25% of the remaining half. The kits were then repackaged for each school session. Four boxes were used per session. Each box contained 7 of the same kit. Thus 4 different kits were used in each session. It took approximately 110 hours to do a quality control on all the kits and repackage them (an average of about 6.5 hours per kit).

Field supervisors sent unused kits back to NCS, the used ones were usually left in the schools where the administration took place. The kits had to be inventoried and stored. It proved very difficult to break out the time spent on these activities from the rest of the receiving activities.

Task Administration

Westat, Inc. used field supervisors to administer the assessment nationally. The time taken to train fifty-five field supervisors on hands-on tasks was factored into the cost of performance assessment. The supervisors were trained in all aspects of assessment administration over a three-day period. Approximately half of this time (12 hours) was spent on becoming familiar with the 17 hands-on tasks. It took about 40 person hours to train on each task.

The time taken by field supervisors to prepare for school sessions increased due to the presence of hands-on tasks. Supervisors had to check the inventory of the boxes of kits, transport them, and return unused kits to NCS. Each supervisor also carried a master kit containing replacement parts. No attempt has been made to factor in these activities as part of the cost of performance assessment.

Each school session typically lasted between 2 and 3 hours. Of this, supervisors spent 20 minutes at grade 4 and 30 minutes at grade 8 and grade 12 administering the hands-on tasks. This time has not been factored in as part of the cost because each task was given in place of a paper-and-pencil block, that would have taken the same time to administer.

Scoring

There were 51 unique blocks of questions field tested (34 paper-and-pencil blocks and 17 hands-on task blocks). In all, 450 questions were scored. The number of constructed-response questions per block is shown in Table 6.

Table 6: Numbers of Constructed-response Questions per Block

GRADE	AVERAGE NUMBER OF CONSTRUCTED-RESPONSE QUESTIONS PER BLOCK	
	Hands-on Task	Paper-and-pencil
4	5	6
8	7	10
12	7	10

It took about 12 hours to train and score one block of items using 1 trainer, 10 scorers, and 1 table leader. This included double scoring and back-reading. Ten percent of the papers were double-scored (Table 7).

Table 7: Average Number of Hours Taken to Assemble Training Packets, Train, and Score One Block of Questions*

ACTIVITY	NUMBER OF PEOPLE	HOURS	PEOPLE X HOURS
Assembly of Training Packets	2	3	6
Training (ETS)	1	3	3
Training of Table Leader and Scorers (NCS)	11	3	33
Scoring** Including 10% Double Scoring	10	9	90
Back Reading	1	9	9
TOTAL			141

* Hands-on task block or paper-and-pencil block
** Assume 1 block x 500 students

It should be noted that constructed-response questions in the latest NAEP assessments are being scored in an image processing system. Student responses are scanned into a computer so that scorers can read student responses on the screen and then enter the scores directly into the computer. A certain percentage of items are read twice and interrater reliability is monitored continually. Back-reading is carried out while scoring is taking place. The image system has resulted in improvements in speed of scoring and reliability.

Summary

Many assessments now contain a variety of question types including multiple-choice and short and long constructed-response exercises. These types of exercises are found in the NAEP science assessment in both paper-and-pencil blocks and hand-on task blocks. It has been the experience of NAEP that all questions take time to develop, especially if they test higher-order thinking skills. Therefore, the time taken to develop a hands-on task block is very similar to the time taken to develop an equivalent set of paper-and-pencil questions. However, additional costs do need to be added for small-scale review, pilot testing, dealing with kit manufacturers, quality control of kits, and repacking.

The training of field supervisors allowed each of them to familiarize themselves with the tasks. Although the training session was conducted at the same time that they were trained to administer the paper-and-pencil blocks, more time was clearly needed to train them to administer the tasks.

The scoring of constructed-response questions in the hands-on task blocks and the paper-and-pencil blocks took place over the same period of time. The time taken to score a hands-on task block was similar to the time taken to score a paper-and-pencil block, despite the fact that scorers attempted each task prior to training.

Table 8 summarizes cost in terms of hours and includes steps common to the development and scoring of a hands-on task block and a paper-and-pencil block.

Table 8: Summary of Cost in Hours/Task

	PROFESSIONAL HOURS	ADMINISTRATIVE ASSISTANCE	TOTAL (HOURS)
*Task Development	69	25	94
QC and Repacking	7		7
Field Supervisor Training	40		40
*Training & Scoring (500 student responses)	141		141
TOTAL	257	25	282

* Includes steps common to a hands-on task block and a paper-and-pencil block

Table 9 summarizes the cost in hours per hands-on task block factoring out the steps that are common to both a hands-on task block and a paper-and-pencil block.

Table 9: Summary of Cost of Hands-on Task Minus Cost for Preparation and Scoring of an Equivalent Paper-and-Pencil Block

	PROFESSIONAL HOURS	ADMINISTRATIVE ASSISTANCE	TOTAL
*Task Development	14	4	18
QC & Repacking	7		7
Field Supervisor Training	40		40
TOTAL	61	4	65

* Unique to development of hands-on task block

The inclusion of hands-on tasks increased the cost in hours/task by approximately 23%. The average cost in hours/500 students is 65 hours. Therefore, the cost per student is 0.13 hours.

In addition to the hours shown above, the cost of kits must also be factored into the analysis. Kits were assembled in-house in order that a small-scale study could occur. Hence, cost was somewhat minimal - approximately $2 per kit. Since approximately 6 of each kit were produced, a total of $12 was spent per hands-on task.

The kits produced by the manufacturers cost on average $6 per kit. Thus the average cost per 500 students was $3,000.

Conclusion

The analysis of the cost of inclusion of hands-on tasks in the NAEP science assessment was complicated by a number of factors. First, a number of organizations interact to produce, administer, and score the NAEP assessment. Second, the machinery for administration is already in place. The replacement of a paper-and-pencil block by a hands-on task block that takes the same amount of time to administer does not add to the cost of the assessment. Third, scoring questions in a hands-on task block takes about the same time as scoring questions in a paper-and-pencil block. Thus, the additional cost per student of inclusion of hands-on tasks in the NAEP science assessment can largely be ascribed to the cost of each kit, and to a lesser extent to small-scale review and pilot testing, field supervisor training, kit quality control, and repacking.